PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 17/60 | A2

(11) International Publication Number: WO 00/42544

(43) International Publication Date: 20 July 2000 (20.07.00)

(21) International Application Number: PCT/US00/01084

(22) International Filing Date: 18 January 2000 (18.01.00)

(30) Priority Data:
09/232,357 15 January 1999 (15.01.99) US

(71) Applicant: IMANDI CORPORATION [US/US]; 14570 NE 95th Street, Redmond, WA 98052 (US).

(72) Inventors: JOHNSON, Eric, W., W.; 16911 NE 106th Street, Redmond, WA 98052 (US). KHER, Raghav, P.; 17436 NE 38th Street, Redmond, WA 98052 (US). JACOBS, Bradley, W.; 29824 – 25th Place South, Federal Way, WA 98003 (US).

(74) Agent: BERGSTROM, Robert, W.; Weiss Jensen Ellis & Howard, Suite 2600, 520 Pike Street, Seattle, WA 98101 (US).

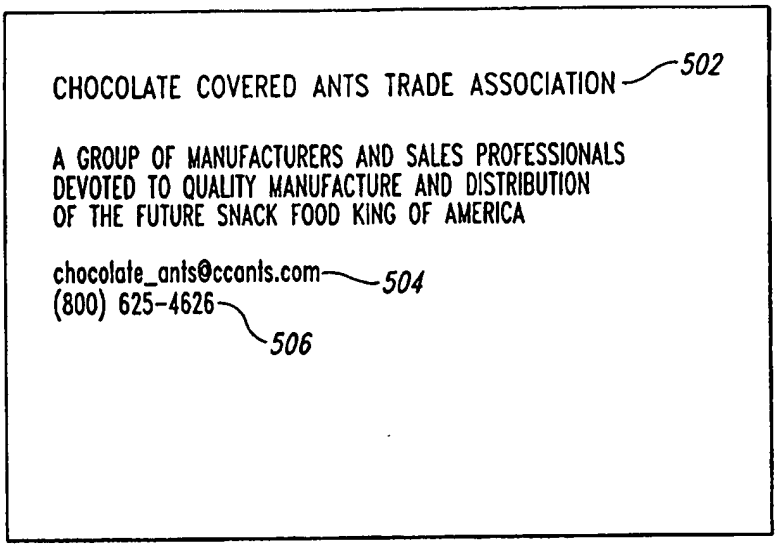
(81) Designated States: AU, BR, CA, CN, IN, JP, KR, NO, NZ, SG, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: EXTRACTION OF VENDOR INFORMATION FROM WEB SITES

(57) Abstract

A database and database creation, maintenance, and update processes and tools for storing vendor information for use in technology–enabled markets. The vendor information stored within the database allows for automated compilations of lists of vendors having an arbitrary geographical proximity to a customer, offering a product or service desired by the customer, and meeting various customer preferences. Database creation and update tools extract information from various information sources, such as Internet–based web sites, and enhance and update the database on a continuous basis.



Extraction of Vendor Information from Web Sites

Technical Field

The present invention relates to technology-enabled retail and wholesale markets, and, in particular, to a method and system for extracting information about merchants from Internet-based web sites and organizing the extracted information to facilitate subsequent searching and retrieval operations needed by technology-enabled retail and wholesale markets.

Background of the Invention

Computer-readable compilations of vendor information are important components of new, Internet-based retail and wholesale markets. Such compilations may be stored as text files or collections of hyper-linked web pages, but are most commonly stored in any of a number of different types of computer database management systems, such as relational database management systems and object-oriented database management systems. Printable and human-readable representations of the information contained in the compilations can be generated automatically by any of a number of different database management system recording tools. The information contained in many vendor databases is also accessible via graphical displays generated in response to input to Internet-based web pages. Examples include the Yellow Pages, http://uswestdex.com, the American Business Information database, http://infousa.com, and the American Society of Travel Agents, http://astanet.com.

In one new type of Internet-based marketplace, in which lists of merchants suitable for submission of requests for quotes by consumers is provided to the consumers based on consumer preferences ,and, where appropriate, on geographical proximity, a database of vendor information having certain specific features and characteristics is most desirable. The database of vendor information needs to be comprehensive and geographically broad-based with regard to vendors, so that any particular consumer has a high likelihood of being matched to several vendors that offer a particular good or service desired by the consumer and that are located

within a reasonable distance from the consumer's residence. The vendor database needs to contain the business name and zip code for each vendor, along with indications of the categories and subcategories of goods and services offered by the vendor. The vendor database must contain sufficient information to uniquely identify each vendor, such as the vendor's address, phone number, and email address. Finally, the vendor database needs to contain, for each vendor, a pointer, or method of contact, for submitting requests for quotes, such as email addresses, online form pointers or links, or fax numbers.

Unfortunately, currently-available databases do not contain all the different types of vendor information needed for the above-mentioned new type of Internet-based market. Many directories of businesses do not contain complete contact information and do not contain detailed information about the goods and services offered by vendors. Other databases that focus on particular types of businesses, such as the American Society of Travel Agents database mentioned above, may contain much of the detailed information required, but lack the depth of content needed by the Internet-based market. A more serious deficiency of many of the currently-available databases is that they require vendors to actively seek to be listed by the organization that compiles and maintains the database. However, in the Internet-based marketplace, it is more desirable to identify vendors from many types of available information, including vendor homepages and web-based references to vendors, so that the broadest possible compilation of vendors can be obtained.

The new, Internet-based market, described above, thus needs a vendor database compiled and organized to include sufficient information for the generation of lists of merchants suitable to particular consumers based on consumer preferences and geographical proximity. The vendor database needs also to be organized to allow for highly automated, continuous update that provides for extracting vendor from a variety of web-based information sources and merging the extracted information into the vendor database. Automated vendor information extraction and database update provide the only cost-effective means for obtaining the breadth of content required by the Internet-based market.

Summary of the Invention

The vendor database component of one embodiment of the present invention is implemented as a relational database. This vendor database component comprises a collection of relational tables that include tables containing detailed vendor information, tables presenting a many-to-many relationship between vendors and categories and subcategories of goods and services offered by the vendors, and a number of tables that facilitate extraction of new vendor information from web-based information sources and merging that extracted vendor information into the above-mentioned vendor information and category tables. The present invention relates both to a vendor database, as well as to a method and system for the highly automated construction, maintenance, and enhancement of the vendor database, including web-based data extraction tools, merging tools, and database construction and update processes.

Brief Description of the Drawings

Figure 1 shows a hypothetical homepage for a manufacturer's association.

Figure 2 represents the displayed list of vendors resulting from entry of the zip code "98104" into the text entry box of the homepage displayed in Figure 1.

Figure 3 depicts the link page referenced by the hyperlink in Figure 1.

Figure 4 represents the web page referenced by the first hyperlink in Figure 3.

Figure 5 represents the web page referenced by the second hyperlink in Figure 3.

Figure 6 represents the web page referenced by the third hyperlink in Figure 3.

Figure 7 is a flow control diagram for the continuous vendor database construction and update processes of the present invention.

Referring to Table 1, there is a category "auto," having category number "436," that is cross-indexed by the SIC number "5511." Within the category "auto" are subcategories related to the retail sale of new and used Fordge and Datsoya automobiles, each having a unique subcategory number. The vendor database category and subcategory classification system may be far more detailed and complex than the SIC classification system, and may be, in part, generated automatically during the data extraction and database update phases of the present invention.

Table 2, below, is a representation of the relational table "Company." This table contains basic information about each identified vendor.

Table 2

Company (simplified)

Company Id	Name	Street Address	City	State	Zip	Email	Phone
36	Big Bad Auto	911 James	Corvalis	OR	97012	bad@bad.com	503 777 1111
94	Jerry's Datsoya	1183 Wood	Smallville	CA	94082	jerrys@p12.com	209 866 7322
113	Elvis Brake	876 Main	Centertown	KA	51032	elvis@kabrake.com	785 934 8106
119	Bob's Copy	Suite 11B 841 First St	Seattle	WA	98104	bobs@copy.com	206 341 1716
136	Eric's Fordge	1101 Greenlake Way	Seattle	WA	98151	eric@tap.com	206 771 1300
147	Anita's Books	710 Baker	Wicksville	NY	20122	anita@wibook.com	315 812 7689
217	Beth's Books	333 Cherry	Zaksville	TN	31261	beth@zak.com	615 548 1542
222	Downtown Ants	600 Pine	Seattle	WA	98104	dta@a.com	206 340 9578

As discussed below in the following subsection, the table "Company" in the preferred embodiment contains many more columns than the simplified version of the table "Company," shown above. However, for the present example, the simplified tables, such as the table "Company," shown above, provide a clear illustration of the nature of data and the interrelationships between the data, stored within the vendor database

Table 10

InterDBI (simplified)

Category Id	Sub category Id	Name	Street Address	City	State	Zip	Email	Phone
8	13	Downtown Ants	520 Pike Street	Seattle	WA	98104	dta@a.com	206 340 9578
8	13	Comer Ants	701 5th Avenue	Seattle	WA	98104	cant@s.com	206 622 4901
8	13	Chocolate Covered Ants Trade Association	701 5th Avenue	Seattle	WA	98104	chocolate_ants@ccants.com	800 825 4626
8	13	Brenda's Chocolate Covered Managerie	3702 State Street	Madison	WI	53708	brendas@madtown.com	608 624 1495
8	11	Brenda's Chocolate Covered Managerie	3702 State Street	Madison	WI	53708	brendas@madtown.com	608 624 1495
8	7	Brenda's Chocolate Covered Managerie	3702 State Street	Madison	WI	53708	brendas@madtown.com	608 624 1495
8	9	Brenda's Chocolate Covered Managerie	3702 State Street	Madison	WI	53708	brendas@madtown.com	608 624 1495

The relational table "InterDB1" has the same columns as the relational table "Company," although any particular entry in the relational table "InterDB1" may not have values for the columns, depending on the completeness of the newly extracted data. For the current example, information extracted and parsed from the vendor list shown in Figure 2 has been entered into the first two rows of the relational table "InterDB1."

Referring back to Figure 1, there is a hyperlink 106 entitled "Links" at the bottom of the manufacturers' association homepage. When the reader positions a cursor over this hyperlink and depresses a mouse button, the reader's browser displays a linkpage with hyperlinks to related web pages. Figure 3 depicts the link page corresponding to the hyperlink in Figure 1. Linkpages may contain tens or hundreds of links to a variey of different types of pages. In the current example, the linkpage contains three links 302, 304, and 306. When the reader positions a cursor over a hyperlink shown in Figure 3, the reader's browser then displays a new web page referenced by the hyperlink. Figure 4 represents the web page referenced by the first hyperlink in Figure 3. Figure 5 represents the web page referenced by the second hyperlink in Figure 3. Figure 6 represents the web page referenced by the third hyperlink in Figure 3.

In general, chains of hyperlinks represent a mathematical graph superimposed over a collection of web pages. Any two nodes, or web pages, within the

graph may be joined by one or more edges, or hyperlinks, and any node, particularly link pages, such as the linkpage shown in Figure 3, may contain numerous hyperlinks directed to other web pages. There are generally many different possible paths by which the graph can be navigated to touch each node, and arbitrarily selected paths may contain redundancies, cycles, and endless loops. Although a human technician may be capable of slowly traversing the nodes within a hyperlink graph, large-scale data extraction from hyperlink graphs is best accomplished by automated means. For IMMM vendor database construction and update, web crawler tools are used to extract information from hyperlink graphs. The web crawler tools start at a given web page and then follow links and parse information displayed on linked web pages in order to identify and extract new vendor information.

Continuing with the current example, once a data extraction tool has extracted the merchant database information by supplying zip codes to homepages, displayed in Figure 1, a web crawler data extraction tool can then be directed to start extracting information from the graph of web pages referenced by the hyperlink 106 on the homepage. The web crawler data extraction tool first navigates to the linkpage displayed in Figure 3, and from that page to each of the web pages displayed in Figures 4-6. When the web crawler tool arrives at the web pages shown in Figure 4, the web crawler tool attempts to identify vendor information displayed on the page. The web crawler tool looks for character strings representative of company names, addresses, phone numbers, email addresses, zip codes, and other information contained in the fields of relational tables "Company" and "InterDB1." In the case of Figure 4, the web crawler data extraction tool determines that no vendor information is displayed on the page. Although some of the words might be construed by the web crawler data extraction tool to represent the name of a vendor, there is not sufficient additional information, such as addresses or phone numbers, within close proximity to any tentative vendor name in order to justify an attempt to extract vendor information from the page.

The web page displayed in Figure 5 represents information of intermediate interest to the web crawler data extraction tool. The web crawler data extraction tool may identify the first line in Figure 5 502 as the name of a vendor and may identify the email address 504 and phone number 506 as the email and phone

number for the vendor. Although the vendor information is incomplete, the web crawler data extraction tool, in this case, may determine that there is sufficient information to enter the extractable data into the relational table "InterDB1" as an incomplete vendor record, shown above in row 3 of Table 10. Finally, when the web crawler data extraction tool arrives at the web pages shown in Figure 6, the web crawler data extraction tool finds sufficient information on the page to construct the tentative vendor database records contained in rows 4-7 of Table 10, above. Note that the web crawler data extraction tool not only identifies vendor information, but also potential business categories and subcategories, thus constructing, in this case, four records corresponding to four different identified business category/subcategory pairs having CategoryId "8" and SubcategoryId's "13," "11," "7," and "9." As can be seen in Table 1, above, these category/subcategory pairs correspond to the following business categories: chocolate covered ants, chocolate covered insects, chocolate covered crustaceans, and chocolate covered echinoderms.

The data extraction tools and web crawler data extraction tools can be run periodically or continuously to build up multiple tables of potential vendor information, such as relational table "InterDB1" shown above as Table 10. These intermediate tables of potential vendor records are then processed by a merging tool. A merging tool analyzes each record, or row, within the intermediate tables in order to either create, based on that row, a new IMMM vendor database record, update an existing IMMM vendor database record, or flag the data contained in the intermediate table row for analysis by a human technician. The merging process can be described with reference to the first record of the relational table "InterDB1," above. First, the merging tool tokenizes the contents of the following columns: "Name," "StreetAddress," and "Phone." As discussed above, these three different pieces of information may be tokenized in different ways. In the case of the contents of the field "Name," tokenization produces the tokens "downtown" and "ants." Tokenization of the contents of the field "StreetAddress" produces the single token "520 pike." Tokenization of the phone number produces the single token "3409578." The merging tool then uses these prepared tokens to find matching entries in the relational tables "NameIndex," "AddressIndex," and "PhoneIndex," displayed above in Tables 7-9. Each

identified match is scored and entered into the relational table "tempdb.dbo.TokenMatch," shown below:

Table 11

tempdb.dbo.TokenMatch

CompanyId	Score
222	20
222	30
222	20

Each row in relational table "tempdb.dbo.TokenMatch" contains the following two fields: (1) "CompanyId," the unique numerical identifier that specifies a vendor company with a matching token entry in Tables 7, 8, or 10; and (2) "Score," an empirical rating of the quality of the match represented by the entry. The three entries in Table 11 correspond to matches of the tokens "downtown" and "ants" with the last two rows of the relational table "NameIndex" and the token "3409578" with the last row of relational table "PhoneIndex." In the present example, all three matches relate to the vendor identified by CompanyId "222." However, in a live, functioning IMMM vendor database, hundreds of matches related to many different companies may be detected and entered into the relational table "tempdb.dbo.TokenMatch.". The merging tool next computes a cumulative total score for each CompanyId having entried in the relational table "tempdb.dbo.TokenMatch" and enters that total score, along with the corresponding CompanyId, into the relational table "tempdb.dbo.TokenSummary," shown below in Table 12:

Table 12

tempdb.dbo.TokenSummary

CompanyId	TotalScore
222	70

Again, in general, relational table "tempdb.dbo.TokenSummary" may contain tens or hundreds of different entries following the analysis of a single record from the relational table "InterDb1" by the merging tool. The merging tool selects the CompanyId or

CompanyIds from the relational table "tempdb.dbo.TokenSummary" associated with the highest total score. In the current example, there is only one entry containing the total score "70" associated with the CompanyId "222." Thus, in the present example, the first record of the relational table "InterDB1" is identified by the merging tool as corresponding to the existing vendor record in the relational table "Company" identified by CompanyId "222."

At this point, the merging tool then does a field-by-field comparison of the information in the selected row of the relational table "InterDB1" with the row in the relational table "Company" identified, via token matching, as describing the vendor represented by the selected row of the relational table "InterDB1." This field-by-field comparison is facilitated by the entry in the relational table "CompanyMerge" having the same value for the field "CompanyId" in the relational table "CompanyMerge" as the value stored in the field "CompanyId" of the identified row of the relational table "Company." Thus, the field-by-field comparison conducted by the merging tool, in the present example, concerns the first row of relational table "InterDB1," the last row in the relational table "Company," and the last row in the relational table "CompanyMerge." First, the merging tool compares the fields "Name" in the relational tables "InterDB1" and "Company." In both tables, the value contained in the field "Name" is identical. Thus, no update is needed. A field-by-field comparison, proceeding through each of the remaining columns in the relational table "InterDB1," reveals only a single difference between the contents of the corresponding fields of the relational tables "Company" and "InterDB1": the value for the field "StreetAddress" in the relational table "Company" is "600 Pine" while the value for the field "StreetAddress" in the relational table "InterDB1" is "520 Pike Street." Thus, the merging tool determines that the StreetAddress for the vendor "Downtown Ants" may have changed. To determine whether or not to update the field "StreetAddress" in relational table "Company," the merging tool consults the relational table "CompanyMerge" to determine the data source and extraction date for the information contained in the field "StreetAddress" in relational table "Company." Then, the merging tool compares the data source to the data source from which the data in the relational table "InterDB1" was extracted, in the case that the data represents data extracted from a single source, or, alternatively, uses the value from a data source

Claims

1. A database, stored in a computer-readable format on a memory device, containing vendor data for a number of previously identified vendors, the database organized for frequent addition of vendor data for newly identified vendors and frequent updating of vendor data for previously identified vendors, the database comprising:

a vendor data component that stores basic vendor data for previously identified vendors, including previously identified vendors' names, addresses, phone numbers, zip codes, email addresses, and fax numbers;

a category mapping component that represents a many-to-many relationship between previously identified vendors and categories of products and services that can be offered by vendors;

a vendor data merge information component that stores an indication of a data source and timestamp for basic vendor data contained in the vendor data component; and

an indexing component that indexes the basic vendor data to allow for rapid identification of previously identified vendors that may be described by newly obtained vendor information.

2. The database of claim 1 implemented within a relational database management system as a collection of relational tables and procedures.

3. The database of claim 2 wherein the vendor data component that stores basic vendor data for previously identified vendors further stores:

vendors' street addresses and a mailing addresses;

vendors' web site addresses;

vendors' secondary phone numbers; and

vendors' toll-free phone numbers.

4. The database of claim 3 wherein the vendor data component that stores basic vendor data for previously identified vendors further stores longitudes and latitudes for vendors.

5. The database of claim 3 wherein the vendor data component comprises a number of relational database tables.

6. The database of claim 2 wherein the category mapping component that represents a many-to-many relationship between previously identified vendors and categories of products and services that can be offered by vendors is implemented as a relational database table that includes columns containing:

identifiers of companies;

identifiers of major business categories for which the vendor offers an item for sale; and

identifiers of business subcategories for which the vendor offers an item for sale.

7. The database of claim 2 wherein the vendor data merge information component that stores an indication of a data source and timestamp for basic vendor data contained in the vendor data component is implemented as a number of relational database tables containing data source and timestamp columns for data columns within the number of relational tables that together compose the vendor data component of the database.

8. The database of claim 2 wherein the indexing component that indexes the basic vendor data comprises a number of relational database tables, each having a column containing identifiers that identify vendors and columns that contain tokens.

9. The database of claim 8 wherein a token represents the results of a tokenizing process that transforms a particular unit of vendor information stored in the vendor data component and that may be specified in a number of different forms into a common token that results from performing the tokenizing process on a number of the different forms that specify the particular unit of vendor information.

10. The database of claim 9 wherein the indexing component includes the following token tables:

an address token table that stores tokenized addresses;

a name token table that stores tokenized names; and

a phone number token table that stores tokenized phone numbers.

11. A method for periodically enhancing a vendor database by extracting vendor information from a digital communications medium, the method comprising:

identifying an information site within the digital communications medium that may contain vendor data for a number of vendors;

determining whether the identified information site contains a desirable quantity of vendor information;

when the identified information site contains a desirable quantity of vendor information,

extracting vendor information from the information site by an automated process, and

placing the extracted information into a temporary storage component; and

merging the extracted vendor information stored in the temporary storage component into the vendor database.

12. The method of claim 11 wherein the digital communications medium is the Internet and wherein the information site is a web page.

13. The method of claim 12 wherein identifying an information site within the digital communications medium that may contain vendor data for a number of vendors further includes analyzing the informational content of the web page, and related web pages, by a human technician.

14. The method of claim 12 wherein identifying an information site within the digital communications medium that may contain vendor data for a number of vendors further includes analyzing the informational content of the web page, and

related web pages, by an automated process implemented as a computer program that parses information from the computer-readable representation of the web page and related web pages.

15. The method of claim 12 wherein determining whether the identified information site contains a desirable quantity of vendor information comprises determining whether the number of vendors for which information can be extracted from the web page, and related web pages, is greater than a threshold value.

16. The method of claim 12 further including, when the identified information site contains a desirable quantity of vendor information, first implementing a computer program to parse vendor information from the computer-readable representation of the web page and related web pages.

17. The method of claim 12 further including, when the identified information site contains vendor information about fewer than a threshold number of vendors,

manually extracting vendor data from the identified information site; and manually merging the manually extracted vendor data into the vendor database.

18. The method of claim 12 wherein, when the information site contains references to additional information sites,
launching an automated navigational data extraction tool to automatically navigate the referenced additional information sites to extract vendor information from the referenced additional information sites and placing the extracted information into the temporary storage component.

19. The method of claim 11 wherein merging the extracted vendor information stored in the temporary storage component into the vendor database further includes:

for each record of vendor information stored in the temporary storage component,

tokenizing a number of fields within the record;

matching the tokenized fields to tokenized vendor data in an indexing component of the vendor database to generate intermediate matches;

summing the generated intermediate matches for each vendor described by data stored in the vendor database in order to produce a score for each vendor described by data stored in the vendor database for which matches are generated;

determining whether the highest produced score represents a definite match, and

when the highest produced score represents a definite match, merging fields within the record into corresponding fields within a record in the vendor database that describes a vendor associated with the highest score.

20. The method of claim 19 wherein determining whether the highest produced score represents a definite match further includes:

determining whether the highest score exceeds an upper threshold, whether the highest score falls between the upper and a lower threshold, or whether the highest score falls below the lower threshold.

21. The method of claim 20 wherein, when the highest score exceeds an upper threshold, the highest score represents a definite match, wherein, when the highest score falls between the upper and a lower threshold, the highest score represents an ambiguity, and wherein, when the highest score falls below the lower threshold, the highest score indicates that the record describes a vendor not described by a record in the vendor database.

22. The method of claim 21 wherein records for which the highest score represents an ambiguity are placed in an ambiguous records storage component for later analysis.

23. The method of claim 21 wherein records for which the highest score indicates that the record describes a vendor not described by a record in the vendor database are used to create new records within the vendor database.